UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

MASTER COPY - DTIC FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

DTIC
ELECTE
MAR 07 1990
S D

AD-A219 387

1b. RESTRICTIVE MARKINGS

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
ARO 23010.6-MA

6a. NAME OF PERFORMING ORGANIZATION	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
Texas A & M Univ.		U. S. Army Research Office

6c. ADDRESS (City, State, and ZIP Code)	7b. ADDRESS (City, State, and ZIP Code)
College Station, TX 77843	P. O. Box 12211 Research Triangle Park, NC 27709-2211

8a. NAME OF FUNDING / SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
U. S. Army Research Office		DAAL03-87-K-0003

8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
P. O. Box 12211 Research Triangle Park, NC 27709-2211				

11. TITLE (Include Security Classification)
Functional Statistical Data Analysis and Modeling

12. PERSONAL AUTHOR(S)
Emanuel Parzen

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 10/1/86 TO 12/31/89	Feb 1990	8

16. SUPPLEMENTARY NOTATION
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Statistical Methods, Statistical Problems, Statistical Computing, Data Analysis, Statistical Science, Statistics

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The goal of this program was to contribute to a unification of statistical methods which applies to the broad diversity of statistical problems (discrete, continuous data; one sample, two sample problems; univariate, multivariate, time series data; parametric, nonparametric, robust, function estimation methods; goodness of fit and probability model identification). It uses statistical computing in an interactive environment to provide ease of use and effective integration of classical and currently emerging styles of statistical data analysis.

(continued on back)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL

90 03 06 050

Statistical science is defined to be the art of analyzing data from as many points of view as possible. A unified framework for statistical reasoning will make it possible to more rigorously draw conclusions by combining the results yielded by different methods which can (and should be) applied to data. This report contains the results of this effort.

ARO 23010.6-M

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

Department of STATISTICS
Statistical Interdisciplinary
Research Laboratory
E415EP@TAMVM1.BITNET

Emanuel Parzen
Distinguished Professor
Phone 409-845-3188
Fax 409-845-3144

FINAL REPORT
February 1990

U. S. ARMY RESEARCH OFFICE

PROJECT DAAL03-87-K0003

"FUNCTIONAL STATISTICAL DATA ANALYSIS AND MODELING"

October 1, 1986–December 31, 1989

PRINCIPAL INVESTIGATOR:
EMANUEL PARZEN
Department of Statistics
Texas A&M University
College Station, TX 77843

Texas A&M Research Foundation
Project No. 5641

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution /
Availability Codes
Dist | Avail and / or Special
A-1

Reproduction in whole, or in part, is permitted for any purpose of the United States Government. This document has been approved for public release and sale; its distribution is unlimited.

SUMMARY OF WORK ACCOMPLISHED

The word "functional" is used to describe statistical methods which involve quantile domain concepts, comparison density functions, and information-entropy-divergence estimation and testing. We call our research program FUNSTAT to communicate that it seeks to be: functional (useful); functional (abstract analysis); functionals (linear functionals of comparison density functions called components); fun; fundamental; and function graphic (graphs should be pictures of functions).

Our research program seeks to contribute to a unification of statistical methods which applies to the broad diversity of statistical problems (discrete, continuous data; one sample, two sample problems; univariate, multivariate, time series data; parametric, nonparametric, robust, function estimation methods; goodness of fit and probability model identification). It uses statistical computing in an interactive environment to provide ease of use and effective integration of classical and currently emerging styles of statistical data analysis.

Statistical science is defined to be the art of analyzing data from as many points of view as possible. A unified framework for statistical reasoning will make it possible to more rigorously draw conclusions by combining the results yielded by different methods which can (and should be) applied to data.

The concept of unification of statistical methods for continuous and discrete data is based on our discovery that many classical statistical methods for goodness of fit (inluding chi-squared tests and their extensior ·-· Read and Cressie (1988) and Rayner and Best (1989)) can be developed in analogous s by expressing them in terms of a comparison density function $d(u)$ for comparing two distributions F and G. For F and G continuous distribution functions with respective probability densities $f(x)$ and $g(x)$, define

$$d(u) = d(u; G, F) = f(G^{-1}(u))/g(G^{-1}(u))$$

For F and G discrete distribution functions with probability mass functions p_F and p_G

respectively, define

$$d(u) = d(u; G, F) = p_F(G^{-1}(u))/p_G(G^{-1}(u))$$

Our paper "Quantile-Information Functional Statistical Inference and Unification of Discrete and Continuous Data Analysis" introduced these concepts.

Quantile spectral analysis introduces the idea of studying long memory time series whose spectral density $f(\omega)$ has zeroes or infinities by studying the rate at which $f(\omega)$ approaches zero or infinity. One approach is to treat the sample spectral density as a data batch and study its long tailed behavior by methods of quantile data analysis. Our paper "Quantile Spectral Analysis and Long Memory Time Series" introduced these concepts.

The study of standard statistical estimators under dependence is an important frontier of modern statistical research. Our work (on distribution of non-parametric two-sample tests computed from stationary time series) provides an outline of how to express the effects of dependence. One develops asymptotic representations of the statistic in terms of sample means of suitable functions of the time series (called influence function representations); the spectral densities at zero frequency of the influence function transformed time series provide the information required to express the effects of dependence. Our paper "Distribution under Dependence of Non-parametric Two-Sample Tests" introduced these concepts.

Our paper "Quantile Statistical Data Analysis" summarizes the basic ideas of the quantile domain approach to analysis of a univariate sample. Reviewed are identification quantile box plot, tail classification of probability laws, identification quantile- quantile plots, cumulative weighted spacings tests of goodness of fit of a univariate probability model, rejection method of simulaton application of comparison density functions.

Extreme value theory, which is becoming increasingly important for applications, can be derived and formulated in a more user friendly way (both for understanding proofs of asymptotic distributions and for providing constructive formulas for norming factors) when developed in terms of quantile functions rather than distribution functions. The quantile

approach was reported in our paper "Quantile Based Unified Distribution of Extreme Values and Order Statistics".

The comparison density estimation approach to unification of nonparametric tests for equality of several samples in introduced in our paper "Multi-sample Functional Statistical Data Analysis".

A framework for the "culture" of statistical practice is proposed in our paper "Statistical Culture: Improving the Practice of Statistics". We propose that statisticians recognize the need to develop maps of statistical methods which will help applied statisticians to strive for continuous improvement of methods, to learn new methods to consider as alternatives, to compare competing methods, to more confidently obtain conclusions from comparison as of results of competing methods of statistical analysis of data of a certain type, to obtain problem-driven results from methods-driven results, to obtain substantive conclusions from data for which prior substantive knowledge was not available. These are among the goals of our overall research program.

This project supported the research as graduate students of Will Alexander and Scott Grimshaw who received Ph.D.'s in 1989; their research is described in their abstracts below.

PUBLICATIONS OF
ARO CONTRACT DAAL03-K0003

Emanuel Parzen, (1985). 'Quantile-Information-Functional Statistical Inference and Unification of Discrete and Continuous Data Analysis,' *Proceedings of the 30th Army Conference on the Design of Experiments*, 213-226.

Emanuel Parzen, (1986). 'Quantile Spectral Analysis and Long Memory Time Series,' *Journal of Applied Probability*, Vol. 23A, 41-55.

Emanuel Parzen, (1986). 'Distribution under Dependence of Nonparameric Two-Sample Tests,' *Proceedings of the 31st Army Conference on the Design of Experiments*, 19-27.

Emanuel Parzen, (1987). 'Quantile Statistical Data Analysis,' *Proceedings of the 32th Army Conference on the Design of Experiments*, 281-292.

Emanuel Parzen, (1987). 'Quantile Based Unified Distribution of Extreme Values and Order Statistics.'

Emanuel Parzen, (1989). 'Multi-Sample Functional Statistical Data Analysis,' *Statistical Data Analysis and Inference Conference in Honor of C. R. Rao*, ed. Y. Dodge, Amstedam: Elsevier.

Emanuel Parzen, (1989) 'Statistical Culture: Improving the Practice of Statistics' Technical Report #62, Department of Statistics, Texas A&M University; *American Statistician.*

Ph.D. Theses

Scott D Grimshaw, (1989) 'A Unified Approach to Estimating Tail Behavior.' Also, Technical Report #55, Department of Statistics, Texas A&M University.

William Pyle Alexander, (1989). 'Boundary Kernel Estimation of the Two Sample Comparison Density Function.' Also, Technical Report #56, Department of Statistics, Texas A&M University.

INVITED LECTURES BY PROFESSOR PARZEN ON FUNCTIONAL STATISTICAL INFERENCE

1986

U. S. Army 32nd Conference on Design of Experiments, Monterey, CA.

1987

University of Chicago

University of Illinois

Harvard University

George Mason University

International Extreme Value Workshop, University of California, Santa Barbara

American Statistical Association National Meeting, San Francisco

U. S. Army 33rd Conference on Design of Experiments, University of Delaware

1988

Brown University

Harvard University School of Public Health

U. S. Army 34th Conference on Design of Experiments, Las Cruces, NM

1989

University of Minnesota

University of California, San Diego

University of California, Los Angeles

University of California, Berkeley

Stanford University

Yale University

New York University

AT&T Bell Laboratories

University of Delaware

Harvard University School of Public Health

University of Minnesota IMA Robustness Workshop

Switzerland Neuchatel Conference in Honor of C. R. Rao

Texas A&M University

University of Texas M. D. Anderson Cancer Center

ASA Houston Chapter (HACASA)

U. S. Army 35th Design of Experiments Conference, Monterey, CA

1990

University of Kentucky

McGill University, Montreal

Interface '90 Symposium on the Interface

American Statistical Association National Meeting Anaheim, CA

Institute of Mathematical Statistics International Meeting, Sweden

American Political Science Association National Meeting, San Francisco

ABSTRACT

Boundary Kernel Estimation

of the Two Sample Comparison Density Function. (May 1989)

William Pyle Alexander, B.S.A., University of Arkansas

Chair of Advisory Committee: Dr. Emanuel Parzen

The focus of this work is to derive functional and graphical statistical techniques for the two sample problem suitable for implementation in modern computing environments. In the two sample problem, it is desired to test the null hypothesis that two independent random samples have a common distribution function. Assuming certain conditions on the distribution functions, a procedure is proposed which has strong graphical elements, a sound theoretical foundation, and estimates the relation of the two distributions if the null hypothesis is rejected. The proposed procedure has as its motivation the estimation of the comparison density and inference concerning its uniformity.

The proposed procedure is both a statistical test of the null hypothesis and a model selection criterion. The test is based on components of a new stochastic process which is termed the kernel density process. This process is based on a boundary kernel estimate of the comparison density. It is proposed to apply a new test, the subset chi-square test, to these components. If the null hypothesis is rejected, the components found to be significant are used to construct a damped orthogonal series estiamte of the comparison density.

The power of the proposed test under local alternatives is compared to two commonly used portmanteau statistics, the Cramér-von Mises and the Anderson-Darling, and to a third statistic suggested by this work. A new method for finding the power of these statistics under local alternatives is given. This method uses the fast Fourier transform to invert an approximation to the characteristic function of the statistic. The proposed test is seen to have good power properties. A simulation study is conducted to examine its small sample size. Its size is found to remain close to its nominal value.

ABSTRACT

A Unified Approach to Estimating Tail Behavior. (May 1989)

Scott D Grimshaw, B.S., Southern Utah State College;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Emanuel Parzen

Tail estimators are proposed which make minimal assumptions and let the data dictate the form of the probability model. These estimators use only the observations in the tail and are based on a unifying density-quantile model. The fundamental result in this work is a representation of the quantile function of the exceedences over a threshold. This representation (1) motivates a unified parameterization for tail estimators of the underlying probability model; (2) motivates methods for obtaining parameter estimates; and (3) simplifies the derivation of the asymptotic properties of the proposed parameter estimates.

Parameter estimates may be obtained using a Generalized Pareto Distribution or a Generalized Extreme Value Distribution model of the exceedences. Assuming the underlying distribution can be correctly classified as either short tailed or long tailed, other estimates are formed. The asymptotic properties of these estimates are derived under rate of convergence conditions to show the effect of threshold selection on parameter properties.

The parameters are shown to be nonidentifiable and their estimators contain a bais which may approach zero very slowly. Therefore, if the parameters are the focus of the analysis, extremely large sample sizes are required to reduce the bias to a negligible amount. If the tail estimates are of interest, the bias is less likely to be serious and the nonidentifiability problem provides a closer approximation to the tail for small sample.